STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN STATISTICAL SCIENCE

END OF SEMESTER EXAMINATIONS

STA 8103: LINEAR MODELS

DATE: Monday 11th January 2023. **TIME: 3 HOURS**

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS.

QUESTION ONE

a. Outline any four approaches you would use for variable selection (8 marks)
b. Non-full rank models are commonplace in linear models, reducing the model to one of full rank is one of the approaches, briefly outline other 2 approaches (6 marks).
c. The following is an output of a linear regression model.

```
Call:
lm(formula = asthma_sx ~ hazards * mutation_present, data = asthma)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4616  -2.9869  -0.1139   3.2278  10.0346

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                48.9800     0.7107  68.914  < 2e-16 ***
hazards                     0.6995     0.2789   2.508  0.01382 *
mutation_presentY           3.6679     1.0259   3.575  0.00055 ***
hazards:mutation_presentY   2.3404     0.4040   5.794 8.72e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.123 on 96 degrees of freedom
Multiple R-squared:  0.5746,    Adjusted R-squared:  0.5613
F-statistic: 43.22 on 3 and 96 DF,  p-value: < 2.2e-16
```

Exhaustively discuss the results. (6 marks)

d. Non-linear least squares regression extends linear least squares regression for use with much larger and more general class of functions. Outline the non-linear least squares optimization using the Gauss-Newton algorithm (6 marks)
e. Outline the local polynomial regression for estimating a function at a point x_0. (4 marks)

QUESTION TWO

Consider the following one-Way ANOVA with 2 groups with three observations each.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}$$

The rank of X is 2. Reducing the model to one of full rank, obtain the estimates for α (15 marks)

QUESTION THREE

a. Regularization techniques in particular ridge and LASSO regression impose a penalty λ to reduce model complexity and prevent overfitting. Outline two approaches that can be used to determine the optimal penalty term (4 marks)
b. Obtain the maximum likelihood estimate for β_0, β_1 *and* σ^2 for a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ (11 marks)

QUESTION FOUR

a. Outline the three components of a generalized linear model (3 marks)
b. Discuss how you would fit a semi-parametric model with a linear spline with one knot (4 marks)
c. Outline any 4 departures from the linear model assumption, their tests and possible remedial measures (8 marks)

QUESTION FIVE

The following is data viral load (Y) and dosage in milligram (X).

Dosage	Viral load
10	48
4	29
14	66
6	33
29	90

i. Test at 5% level the significance of the intercept and drug dosage (8 marks)
ii. What percentage of the variation in Viral load is explained by Dosage (3 marks)
iii. Obtain the 95% confidence interval for β_0 (4 marks)